On the Prediction of Isolation, Release, and Decease for COVID-19 Patients: A Case Study in South Korea

Tarik Alafif[a,*], Reem Alotaibi[b], Ayman Albassam[a], Abdulelah Almudhayyani[a]

[a]*Computer Science Department, Jamoum University College, Umm Al-Qura University, Jamoum, Saudi Arabia*
[b]*Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia*

Abstract

A respiratory syndrome COVID-19 pandemic has become a serious public health issue nowadays. The COVID-19 virus has been affecting more than twenty-two million people worldwide. Some of them have recovered and have been released. Others have been isolated and few others have been unfortunately deceased. In this paper, we apply and compare different machine learning approaches such as decision tree models, random forest, and multinomial logistic regression to predict isolation, release, and decease states for COVID-19 patients in South Korea. The prediction can help health providers and decision makers to distinguish the states of infected patients based on their features in early intervention to take an action either by releasing or isolating the patient after the infection. The proposed approaches are evaluated using Data Science for COVID-19 (DS4C) dataset. An analysis of DS4C dataset is also provided. Experimental results and evaluation show that multinomial logistic regression outperforms other approaches with 95% in a state prediction accuracy and a weighted average F1-score of 95%.

Keywords: COVID-19, prediction, classification, decision tree, random forest, multinomial logistic regression.

1. Introduction

The COVID-19 pandemic has first appeared in Wuhan, Hubei province, China in December 2019. It has become a serious public health issue worldwide as it spreads quickly. The COVID-19 virus is an infectious disease which directly affects people lungs. It is a branch of Coronaviruses family. It is believed that it has been transmitted from animals to human.

*Corresponding author.
Email addresses: tkafif@uqu.edu.sa (Tarik Alafif), ralotibi@kau.edu.sa (Reem Alotaibi), asbassam@uqu.edu.sa (Ayman Albassam), s44181770@st.uqu.edu.sa (Abdulelah Almudhayyani)

The virus also transmits from human respiratory to another which is already noticed worldwide. The virus may cause mild or severe symptoms to affected people during the virus replications in their bodies. Affected people may develop symptoms such as dyspnea, fever, sore throat, cough, fatigue, and pneumonia. The virus has become a serious disease since it may cause death after developing the symptoms. Fortunately, many people have recovered and have been released while others are still isolated. While many people are recovered, the number of deaths is very few compared to the number of recovered ones.

Machine learning (ML) algorithms play an important role in today's research because of the ability to automate complex tasks. These algorithms can learn from previous experience to predict future outcomes. Many research works have been developed and used in machine learning. Some of the well-known ML classifiers are Decision Trees (DTs), Random Forest (RF), and Multinomial Logistic Regression (MLR).

DT, so-called CART, is a popular supervised machine learning and a top-down predictive modelling approach [1]. It uses a simple tree structure representation to classify negative and positive examples recursively based on their input features. Various decision tree approaches have been proposed over the past years. ID3 [2], C4.5 [3] and CART [4] are the most common ones [5].

RF is an ensemble learning approach that is based on generating a large number of decision trees using a different subset of the training data [6, 7]. These subsets are chosen by random sampling of the original training data. The final predictions are made by taking the majority vote from all individual classification trees. RF is a powerful algorithm in machine learning. One of the advantages of RF is the ability to handle large datasets with high-dimensional feature space. It can handle thousands of input variables and identify the most significant variables.

In addition to DT and RF, regression approaches are also commonly used to predict future events in many research fields. In the machine learning context, logistic regression is a well-known technique that predicts binary outcomes. On the other hand, MLR is an extended version of a binary logistic regression approach that allows performing multi-class prediction where prediction outcome can be either dichotomous or continuous [8].

Our work is motivated by the success of CARTs, RF, and MLR approaches in many predictive applications such as in vegetation distributions [9], microRNA precursors [10], and response variables [11]. In this paper we propose to use them to predict isolation, release, and decease states for COVID-19 patients in South Korea. The proposed approaches are evaluated using Data Science for COVID-19 (DS4C) dataset. An analysis for DS4C dataset is provided. Experimental results and evaluation show that MLR outperforms other approaches with 95% in a state prediction accuracy and a weighted average F1-score of 95%.

The remainder of this paper is organized as follows. In **Section 2**, related work is reviewed. **Section 3** introduces basic notations and mathematical models used throughout the paper. In **Section 4**, we briefly describe and analyze DS4C dataset. In **Section 5**, we present our work. In **Section 6**, experimental results and evaluation are provided using public DS4C

dataset. We provide discussion details in **Section 7**. Finally, our conclusion and future work are provided in **Section 8**.

2. Related Work

Many research works have been published recently to analyze and tackle the problem of COVID-19 virus in many fields.

Several methods were experimented in [12–24] for COVID-19 analysis and predictions. Fanelli and Piazza [12] analyzed and forecasted COVID-19 spreading in China, Italy, and France. Nadia and Hazem [13] studied the effects of sex, region, infection reason, birth year, release date, and the diseased date on the recovered and deceased cases for COVID-19 patients. Hu et al. [14] applied a stacked autoencoder and K-means to model the transmission dynamics of the epidemics based on provinces and cities. The model was built to forecast the confirmed cases of COVID-19. Remuzzi and Remuzzi [15] predicted the number of infected patients in Italy using an exponential curve. Tartai and Varallyay [16] attempted to predict the results of COVID-19 epidemic in a region using a logistic model. Pedersen and Meneghin [17] analyzed and predicted the spread of the COVID-19 virus using restrictions to reduce the epidemic in Italy. Peng et al. [18] proposed a SEIR model to analyze the epidemic in China. Petropoulos and Makridakis [19] forecasted the confirmed cases, death cases, and recovery cases of COVID-19 using time series forecasting approaches. Chatterjee et al. [20] and Jia et al. [21] developed mathematical models to study the impact of COVID-19 epidemic. Weissman et al. [22] proposed a non-learning predictive mathematical model, called SIR, to predict hospital capacity needs during the COVID-19 pandemic. However, the SIR model predicts only the epidemic spread over time for the whole population within a specific area. Yang et al. [23] used a long term short memory neural network to predict the COVID-19 epidemic in China. Similar to [23], Vattay [24] attempted to predict the epidemic in Italy.

Different from the aforementioned research works to tackle the COVID-19 problem from different perspectives, we apply and compare different machine learning approaches such as CARTS, RF, and MLR to predict isolation, release, and decease states for COVID-19 patients in South Korea. The proposed approaches are evaluated using DS4C dataset to measure the performance of the predictions.

3. Preliminaries

Let D a set of patients records such that $D = \{X_1, X_2, \ldots, X_n\}$, where n is the number of patients. Each patient record X is belonging to one of the classes j that is used for the classification task such that $D = \{(X_1, j_1), (X_2, j_2), \ldots, (X_n, Y_n)\}$. Each patient record X is composed of several input features such that $X = \{x_1, x_2, \ldots, x_m\}$, where m is the total number of patient features. Then, each feature becomes a predictor variable of the class j such that $X = \{(x_1, x_2, \ldots, x_m, j)\}$, Y is the class variable taking values either in a set of classes $C = \{1, 2, \ldots, k\}$, where k is the total number of classes.

Next, we show the basic mathematical equations used in the following approaches:

3.1. CARTs

In CARTs, Gini impurity is used to measure the purity in the decision tree models by knowing how good the split of input features in each node [25, 26]. Then, the Gini impurity is computed using the training examples with the class j as shown in Equation 1:

$$\text{Gini}(D|x) = 1 - \sum_{j=1}^{k} {P_j}^2 \tag{1}$$

where x, k and P_j are the examined feature, the number of classes and the probability of the j^{th} class respectively.

The Gini Index is computed for each input feature partition and the average Gini Index for the examined feature x is defined as:

$$\text{Gini}(D|x) = \sum_{i=1}^{vals(x)} \frac{|D_i|}{|D|} \text{Gini}(D_i|x) \tag{2}$$

where x, D_i and $vals(x)$ are the examined feature, a data partition and number of discrete values in feature x respectively. The feature with the lowest Gini Index is chosen as the splitting feature.

3.2. RF

An RF consists of a collection of decision trees $T = \{1, \ldots, ntree\}$. Each decision tree is trained on a different subset of the data using bootstrap sampling (bagging) out of D. Final predictions are made using the majority votes. The RF employs random feature selection for each node of every tree in the forest, meaning that the splitting feature x is chosen at random.

3.3. MLR

Logistic regression is defined as a generalized linear model that is usually used to classify binary outcomes [8]. Ridge estimator is proposed to improve model prediction [27]. Logistic regression approach is slightly modified to deal with a multi-class classification that is called the MLR, in case of predicting k outcome classes for n instances with m observations, the parameter matrix of regression coefficients B to be calculated will be an $m * (k - 1)$ matrix. e is the exponential function. Class j probability with the exception of the last class is shown in Equation 3:

$$P_j(X_i) = \frac{e^{X_i B_j}}{\sum_{j=1}^{k-1} e^{X_i B_j} + 1} \tag{3}$$

The last class probability is shown in Equation 4:

$$1 - \sum_{j=1}^{k-1} P_j X_i = \frac{1}{\sum_{j=1}^{k-1} e^{X_i B_j} + 1} \tag{4}$$

The (negative) multinomial log-likelihood is shown in Equation 5::

$$L = -\sum_{j=1}^{n} \{ \sum_{j=1}^{k-1} Y_{ij} * ln(P_j(X_i) + (1 - (\sum_{j=1}^{k-1} Y_{ij})) * ln(1 - \sum_{j=1}^{k-1} P_j(X_i))\} + ridge * (B^2) \tag{5}$$

3.4. Evaluation Metrics

The accuracy for multi-class classification is defined as the number of instances correctly

identified as either truly positive or truly negative out of the total as shown in Equation 6. The error rate is the complement of accuracy.

$$\text{accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{6}$$

where TP, TN, FP and FN represent true positive, true negative, false positive and false negative respectively.

F1-score is averaging the harmonic mean of precision and recall as shown in Equations 7:

$$\begin{aligned} \text{precision} &= \frac{TP}{TP + FP} \\ \text{recall} &= \frac{TP}{TP + FN} \\ \text{F1-score} &= \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}} \end{aligned} \tag{7}$$

where precision and recall represent precision and recall respectively. The weighted average F1-score is computed by averaging F1-score class-by-class while weights are computed based on each target class frequency as shown in Equation 8. Using such evaluation metric helps to keep an account for imbalance distribution of classes when computing the F1-score.

$$\text{Weighted average F1-score} = \frac{\sum_{j=1}^{k} w_j * F1\text{-}score_j}{\sum_{j=1}^{k} w_j} \tag{8}$$

Where k is the total number of target classes, w_j the weighting parameter of the j^{th} class and F1-score$_j$ is the computed F1-score of the j^{th} class.

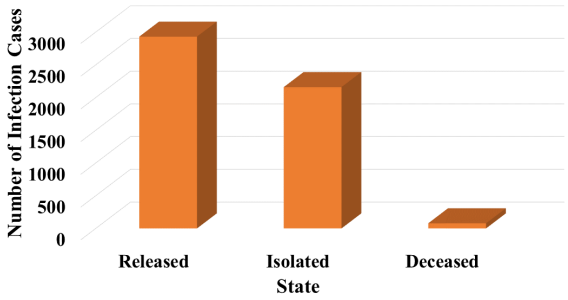

Figure 1: The number of samples distribution for each state in DS4C dataset.

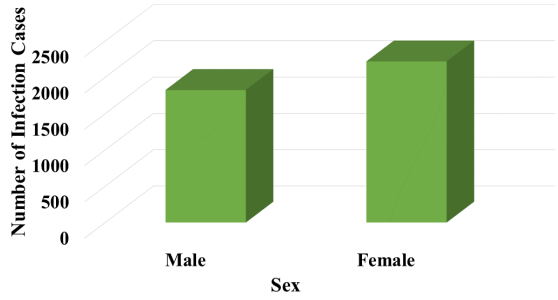

Figure 2: The samples distribution for sex feature in DS4C dataset.

4. DS4C Dataset

DS4C [28] has been published on Kaggle recently into public in February 24th, 2020 in South Korea [1]. The dataset consists of different data in four sheets (Case data, patient data, time-series data, and additional data). The case data describes the data for COVID-19 infectious cases. The patient data describes the epidemi-

[1] https://www.kaggle.com/kimjihoo/coronavirusdataset.

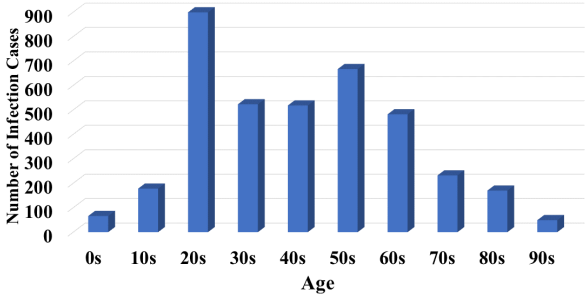

Figure 3: The samples distribution for age feature in DS4C dataset.

ological and route data of COVID-19 patients. The time-series data describes the status of time-series data of COVID-19 patients. Also, additional data is reported for regions, weather, and population. Each sheet is found in the form of a CSV file. In our work, we only use the patient data file to reach the aim of our research since it contains epidemiological data for COVID-19 patients in South Korea.

Patients data consists of 5,165 patients labeled records. Each patient record consists of several features such as sex, age, country, province, city, infection_case, infected_by, contact_number, symptom_onset_date, confirmed_date, released_date, deceased_date, and state. The state feature represents the label for each patient record. The dataset has three states. Each record in the dataset is either labeled isolated, released, or deceased.

Figure 1 shows the states versus the number of infection cases in the dataset. The figure also shows clearly the unbalanced number of the sample's distribution in this dataset which consists of 2,158 isolated states, 2,929 released states, and 78 deceased states. We also show the samples distribution for sex and age features in the dataset in Figure 2 and Figure 3 respectively. We notice that the number of infected male is closed to the number of infected female. We also notice that the most infected age among patients is between 20 to 29 years old.

We have investigated the causes of COVID-19 infection in South Korea patients in this dataset. We have found the most patients receive this infection are from the most frequent causes "Contact patient". Then, "Overseas inflow" is followed. Figure 4 depicts clearly the most common causes of COVID-19 infection in South Korea patients. Therefore, a social distancing is needed to halt or at least decrease the spread of the COVID-19 pandemic.

Figure 4: The causes for COVID-19 infection in South Korea according to the infection cases in DS4C dataset.

5. Our Work

We propose to use DT models, RF, and MLR to classify the states (isolated, released,

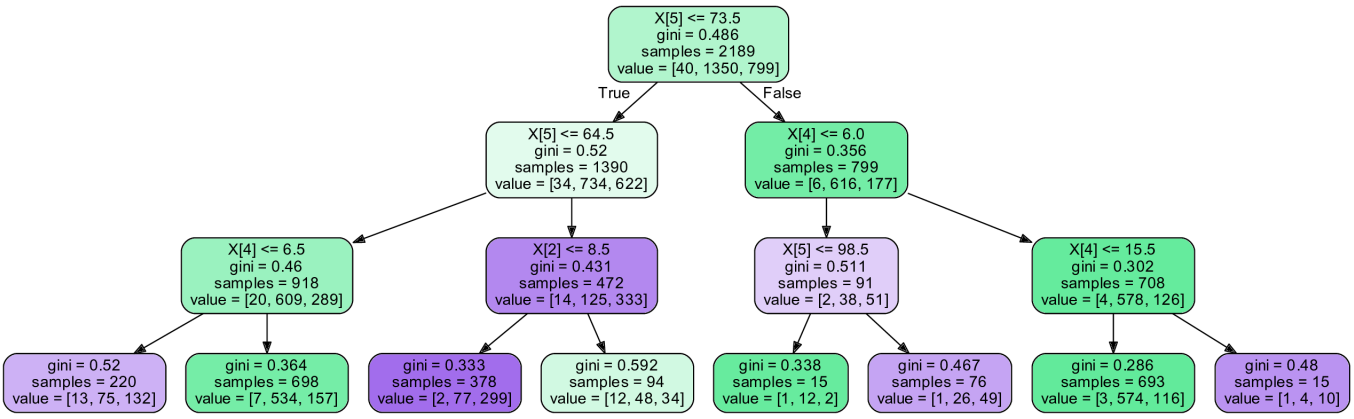

Figure 5: The generated DT architecture of our approach using the maximum DT depth of 3.

and deceased) for COVID-19 patients in South Korea. The DT models learn discriminative features from patient records to perform the prediction. The models predict the patient's state based on several feature variables. We train our DT models using the features such as sex, age, country, province, city, infection_case, contact_number, symptom_onset_date, confirmed_date, released_date, deceased_date, and state. Patients case_id are excluded. Before training the models, we pre-process the patients' records to manipulate missing data and non-integer values. The missing data are replaced by -1.

We apply different maximum DT depths to build our predictive intelligent models. The models are trained using the patients' records. For DT models, we use maximum DT depths of 3, 5, and 10 in our work. Figure 5 shows the generated DT architecture after training using the depth of 3. The DT architectures with the maximum depth of 5 and 10 generate a very large tree which consists of many branches and many leaf nodes and non-leaf nodes. A generalization capability is lost if the depth is more than 10. For the RF model, the model is trained with the default parameters according to randomForest package in R [6]. The default parameters in the randomForest algorithm: *ntree*, the number of classification tree (the default value is 500) and *mtry*, the number of features tested at each node (the default values is $sqrt(m)$, where m is the total number of features).

For the MLR model, the model-building process is carried out using weka tool with a ridge estimator which is considered as a stable implementation. The predictive model is fed with a full set of features. The Quasi-Newton method is used to perform the optimization task. Ridge values of 1×10^8 are used in the log-likelihood calculation.

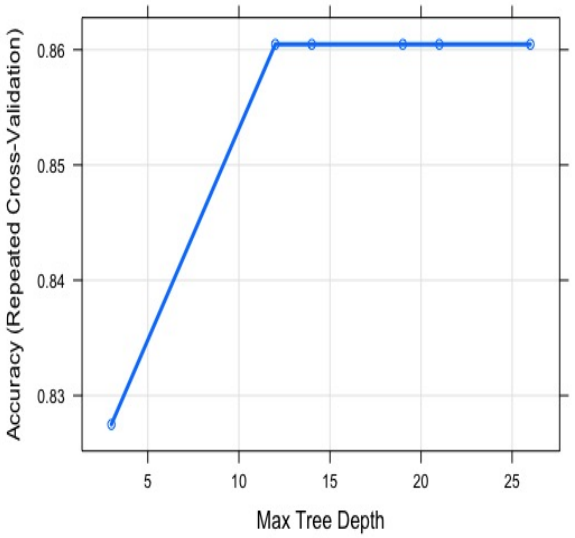

Figure 6: Maximum tree depth tuning based DT.

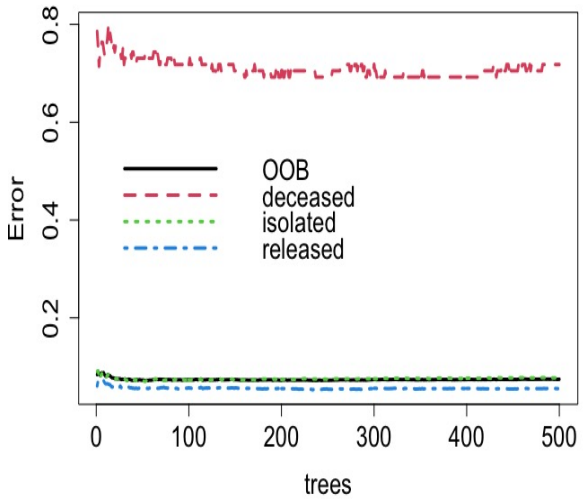

Figure 7: RF error rates with an increase in the number of classification trees. The number of trees and the estimated error rates are shown on x-axis and y-axis, respectively.

6. Experimental Results and Evaluation

Tree-based approaches are implemented in R using *rpart* and *randomForest* packages [6, 29] while MLR approach was built using the Weka version 3.8.4 of machine learning tool [30]. A single laptop 1.6 GHz Dual-Core Intel Core i5 CPU with 8 GB of RAM is used to train and test the models. 10-fold cross-validation is applied for training and testing the models using DS4C dataset. A seed state is set to 1234.

Table 1 shows the accuracy, error rates, and the weighted average F1-score for the tested models. For DT models, we choose three different tree depths: 3, 5, and 10. In order to obtain the effect of the maximum tree depth parameters on the experimental results, we show the performance curve to verify the effectiveness of depth parameters on the models. Figure 6 shows the models accuracy versus different tree depths. It clearly shows that the DT models accuracy stops improving at the maximum depth of 12.

Figure 7 shows the effect of the number of generated trees in the RF on the classification error rate. We can see that the error rate becomes stable when the number of trees is increased. The curve in black color represents out-of-bag (OOB) error rate and the other colors represent misclassification error rate curves. The out-of-bag error is estimated internally during building the DTs.

The confusion matrices using 10-fold cross-validation are provided in Table 2. The table shows true positives, true negatives, false positives, and false negatives test examples after applying each model. We notice that the DT models using different depths are unable to predict the true positives in the deceased state. This

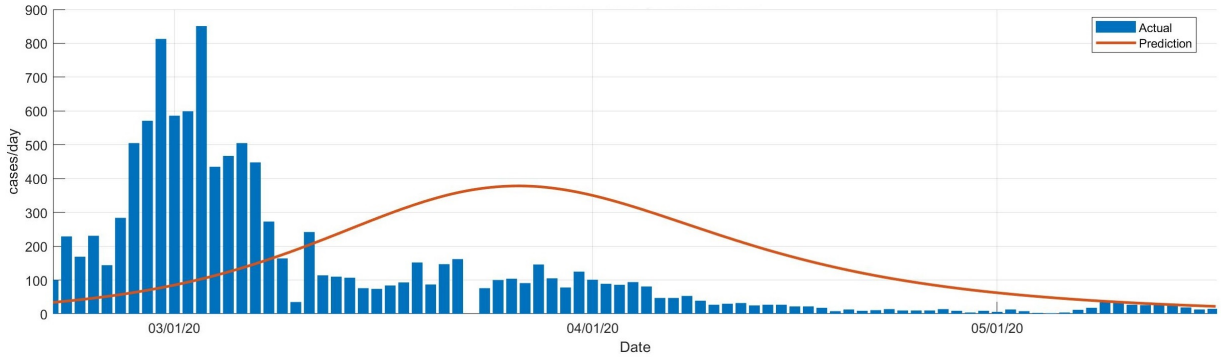

Figure 8: Predictive and actual infection rates by the SIR model in South Korea from March 1, 2020 until May 6, 2020.

Table 1: Prediction accuracy, error rates and the weighted average F1-scores for the applied algorithms using DS4C dataset.

Algorithms implemented	***Accuracy***	***Error rate***	***Weighted average F1-score***
DT (Depth = 3)	82.92%	17.08%	81.74%
DT (Depth = 5)	85.63%	14.37%	84.67%
DT (Depth = 10)	88.21%	11.79 %	87.47%
RF	92.55%	07.45%	92.28%
MLR	**95.00%**	**05.00%**	**95.00%**

maybe due to the small number of examples found in the deceased state. Following the DT models, the RF approach is able to classify 22 true positive examples from the deceased state. Lastly, the MLR model classifies 44 true positive examples from the deceased state correctly. Besides, the small number of false negatives reveals an encouraging result. From Table 1, the MLR approach outperforms other approaches with 95% in the state prediction accuracy and with a margin of 2.45% to RF approach. To the best of our knowledge, there is no existing similar work to compare with ours since the COVID-19 research area and the DS4C dataset are new.

7. Discussion

One important outcome of our study is that MLR model provides higher prediction results than DT and RF classification models when using the COVID-19 DS4C dataset. The reason is that logistic regression model tends to produce more reliable classification outcomes when it deals with few numbers of features, and big amounts of training examples which is the case in the utilized dataset. Therefore, MLR provides most significant performance in prediction accuracy and the weighted average F1-score. On the other hand, RF model performance rel-

Table 2: Confusion matrices for actual versus predicted patients' states.

DT (Depth=3)			
	Deceased	***Isolated***	***Released***
Deceased	0	0	0
Isolated	21	1444	90
Released	57	714	2839
DT (Depth=5)			
	Deceased	***Isolated***	***Released***
Deceased	0	0	0
Isolated	23	1586	92
Released	55	572	2837
DT (Depth=10)			
	Deceased	***Isolated***	***Released***
Deceased	0	0	0
Isolated	25	1808	181
Released	53	350	2748
RF			
	Deceased	***Isolated***	***Released***
Deceased	22	18	38
Isolated	0	1992	166
Released	0	163	2766
MLR			
	Deceased	***Isolated***	***Released***
Deceased	44	14	20
Isolated	15	2042	101
Released	19	89	2821

atively competes with the MLR model performance with a margin of 2.45% in state prediction accuracy. In general, based on the characteristics of features, and corresponding experimental results using DS4C dataset, we can conclude that the MLR model is more effective than other classification models for predicting COVID-19 patients states.

Unfortunately, none of the existing COIVID-19 public datasets provide such adequate features and labeling for predicting the states of isolation, recovery, and decease. Most of the current public COVID-19 datasets focus on providing a statistical data about the death number, infection number, infected areas, patients tweets, infected X-ray scans, and infected CT scans from chests and brains. Also, the current DS4C public dataset lacks some of important features associated with the impact of the COVID-19 virus such as patients' vital signs, medications usage, and chronic diseases. The lack of COVID-19 data and the mechanism of its collection and collaboration is considered a limitation, and a challenge in COVID-19 research area.

8. Conclusion

This research attempts to apply and compare different machine learning approaches such as DT models, RF, and MLR to predict isolation, release, and decease states for COVID-19 patients in South Korea. The proposed approaches are evaluated using DS4C dataset. An analysis for DS4C dataset is also provided. Experimental results and evaluation show that MLR outperforms other approaches with 95% in the state prediction accuracy and the weighted average F1-score of 95%. The proposed machine learning models can help health providers and decision makers in early intervention to take an action either by releasing or isolating the patient after the infection.

Based on our observations from South Korean public COVID-19 statistical dashboard of daily confirmed cases, it reveals that South Ko-

rea has the highest number of infection in February, 2020 with 909 confirmed cases. Figure 8 shows the the predictive and actual infection rates by the SIR model [31] in South Korea from March 1, 2020 until May 6, 2020. Luckily, the current infection number per day has lowered to less than 50 since the beginning of March, 2020. Still, South Korea has a less number of infection rate compared to other countries which may be due to social distancing.

In the future, we plan to collect clinical data from local hospitals to analyze more important features and explore their patterns. These patterns may give a better understanding to the relations of the states. They are maybe beneficial to reduce the impact of the decease states. We also plan to use deep learning based methods such as deep neural networks to predict the state more accurately.

Declarations of interest

None

Acknowledgment

This work is funded by Research and Development Grants Program for National Research Institutions and Centers (GRANTS), Target Research Program, Infectious Disease Research Grant Program, King Abdulaziz City for Science and Technology (KACST), Kingdom of Saudi Arabia, grant number (5-20-01-007-0001).

References

[1] L. Rokach, O. Z. Maimon, Data mining with decision trees: theory and applications, volume 69, World scientific, 2008.

[2] R. Quinlan, Induction of decision trees, Machine Learning 1 (1986) 81–106. URL: http://dx.doi.org/10.1023/A:1022643204877. doi:10.1023/A:1022643204877.

[3] R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[4] L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone, Classification and Regression Trees, Chapman and Hall, New York, 1984.

[5] A. Priyam, G. Abhijeeta, A. Rathee, S. Srivastava, Comparative analysis of decision tree classification algorithms, International Journal of current engineering and technology 3 (2013) 334–337.

[6] A. Liaw, M. Wiener, Classification and regression by randomforest, R News 2 (2002) 18–22. URL: https://CRAN.R-project.org/doc/Rnews/.

[7] T. K. Ho, Random decision forests, in: Third International Conference on Document Analysis and Recognition, ICDAR 1995, August 14 - 15, 1995, Montreal, Canada. Volume I, IEEE Computer Society, 1995, pp. 278–282. URL: https://doi.org/10.1109/ICDAR.1995.598994. doi:10.1109/ICDAR.1995.598994.

[8] D. W. Hosmer, Applied logistic regression, Wiley series in probability and statistics, 3rd ed. ed., Wiley, Hoboken, N.J, 2013.

[9] D. Moore, B. Lees, S. Davey, A new method for predicting vegetation distributions using decision tree analysis in a geographic information system, Environmental Management 15 (1991) 59–71.

[10] P. Jiang, H. Wu, W. Wang, W. Ma, X. Sun, Z. Lu, Mipred: classification of real and pseudo microrna precursors using random forest prediction model with combined features, Nucleic acids research 35 (2007) W339–W344.

[11] L. Breiman, J. H. Friedman, Predicting multivariate responses in multiple linear regression, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 59 (1997) 3–54.

[12] D. Fanelli, F. Piazza, Analysis and forecast of covid-19 spreading in china, italy and france, Chaos, Solitons & Fractals 134 (2020) 109761.

[13] A.-R. Nadia, A.-N. Hazem, Data analysis of coronavirus covid-19 epidemic in south korea based on recovered and death cases, Journal of Medical Virology (2020).

[14] Z. Hu, Q. Ge, L. Jin, M. Xiong, Artificial intelligence forecasting of covid-19 in china, arXiv preprint arXiv:2002.07112 (2020).

[15] A. Remuzzi, G. Remuzzi, Covid-19 and italy: what next?, The Lancet (2020).

[16] D. Tátrai, Z. Várallyay, Covid-19 epidemic outcome predictions based on logistic fitting and estimation of its reliability, arXiv preprint arXiv:2003.14160 (2020).

[17] M. G. Pedersen, M. Meneghini, Quantifying undetected covid-19 cases and effects of containment measures in italy, ResearchGate Preprint (online 21 March 2020) DOI 10 (2020).

[18] L. Peng, W. Yang, D. Zhang, C. Zhuge, L. Hong, Epidemic analysis of covid-19 in china by dynamical modeling, arXiv preprint arXiv:2002.06563 (2020).

[19] F. Petropoulos, S. Makridakis, Forecasting the novel coronavirus covid-19, PloS one 15 (2020) e0231236.

[20] K. Chatterjee, K. Chatterjee, A. Kumar, S. Shankar, Healthcare impact of covid-19 epidemic in india: A stochastic mathematical model, Medical Journal Armed Forces India (2020).

[21] L. Jia, K. Li, Y. Jiang, X. Guo, et al., Prediction and analysis of coronavirus disease 2019, arXiv preprint arXiv:2003.05447 (2020).

[22] G. E. Weissman, A. Crane-Droesch, C. Chivers, T. Luong, A. Hanish, M. Z. Levy, J. Lubken, M. Becker, M. E. Draugelis, G. L. Anesi, et al., Locally informed simulation to predict hospital capacity needs during the covid-19 pandemic, Annals of internal medicine (2020).

[23] Z. Yang, Z. Zeng, K. Wang, S.-S. Wong, W. Liang, M. Zanin, P. Liu, X. Cao, Z. Gao, Z. Mai, et al., Modified seir and ai prediction of the epidemics trend of covid-19 in china under public health interventions, Journal of Thoracic Disease 12 (2020) 165.

[24] G. Vattay, Predicting the ultimate outcome of the covid-19 outbreak in italy, arXiv preprint arXiv:2003.07912 (2020).

[25] L. Rokach, O. Maimon, Data Mining with Decision Trees: Theory and Applications, World Scientific Publishing Co., Inc., River Edge, NJ, USA, 2008.

[26] L. E. Raileanu, K. Stoffel, Theoretical comparison between the gini index and information gain criteria, Annals of Mathematics and Artificial Intelligence 41 (2004) 77–93. URL: http://dx.doi.org/10.1023/B:AMAI.0000018580.96245.c6. doi:10.1023/B:AMAI.0000018580.96245.c6.

[27] S. le Cessie, J. van Houwelingen, Ridge estimators in logistic regression, Applied Statistics 41 (1992) 191–201.

[28] Datartist, Data Science for COVID-19 (DS4C), https://www.kaggle.com/kimjihoo/coronavirusdataset, 2020. [Online; accessed 25-July-2020].

[29] T. Therneau, B. Atkinson, B. Ripley, Rpart: Recursive partitioning and regression trees. r package version 4.1-8, R News (2014). URL: http://CRAN.R-project.org/package=rpart.

[30] E. Frank, M. A. Hall, I. H. Witten, The WEKA workbench, in: Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", fourth edition ed., Morgan Kaufmann, 2016.

[31] W. O. Kermack, A. G. McKendrick, A contribution to the mathematical theory of epidemics, Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character 115 (1927) 700–721.